



---

# Machine Learning prediction, Econometric modelisation and Economic analysis applied to Cardiovascular diseases

---

*Authors:*

Barbey Charlotte - Domergue Corentin - Prutki Lucas

Sorbonne School of Economics,  
University of Pantheon-Sorbonne.

Email: [lucas.prutki@etu.univ-paris1.fr](mailto:lucas.prutki@etu.univ-paris1.fr)

January 2021

## Abstract

Can statistical modelling, through estimation and prediction, serve as a decision criterion to ensure effective medical prevention? Giving a cost to health, to choose the most cost-effective treatment is very complicated. Our medical economic analysis first seeks to identify the most important risk factors in the development of cardiovascular diseases. Then, in a second phase, to know how and to whom should be the prevention of these diseases.

**Keywords:** Data Visualization, Statistics Tests, Logistic Regression, Machine Learning, F-measures, Confusion Matrix, Quality-Adjusted Life Year (QALYs), Incremental Cost Effectiveness Ratio (ICER), Framingham Score Point.

# 1 Introduction

Cardiovascular diseases hold the sad record of leading cause of death in the world. The World Health Organization estimates that every year, no one dies more from these diseases than any other cause. This affects 17.7 million deaths from these diseases, or 31 percents of total global mortality. It is also observed that more than 75 percents of these millions of deaths occur in low- and middle-income countries, where current health spending is very low and the development of the health system as well.

More generally, what is cardiovascular disease? It is a set of disorders affecting the blood vessels that feed the muscles, the brain, the limbs of the human body but also the heart. Heart attacks and strokes (strokes) are generally the most common and related to the obstruction of an artery that irrigates the heart or brain with blood. Their cause is explained by numerous risks – detailed later – that will increase the possibility of developing these diseases. What are the risk factors? Is prevention before having cardiovascular disease cost-effective? Our research work is part of a preventive framework, it is possible to prevent most cardiovascular diseases by looking at risk factors such as smoking, excessive alcohol consumption, sedentary lifestyles, poor diet, hypertension, hyperlipidemia or diabetes. Early detection of these risk factors could make it more cost-effective, particularly in the treatment of these diseases, whether through better nutrition or medication.

In order to answer the question and try to show that prevention on individuals who have not yet developed cardiovascular diseases is cost-effective, we used a dataset available on *Kaggle* (url: Cardiovascular Disease dataset | Kaggle). This database includes 70,000 patient observations (no indication of the country is provided), along with many variables that will serve as risk factors to focus on for early detection of these diseases.

Variables List	
Variables	Informations
Id	It is an identifier assigned to each patient.
Age	This is the age of each patient in days.
Gender	Binary variable indicating the gender of each individual (Male = 2 or Female = 1).
Height	This is the height of the patient (in Cm).
Weight	This is the weight of the patient (in Kg).
Ap hi	Systolic blood pressure. During the systolic phase (in Mmhg).
Ap lo	Diastolic Blood Pressure. During the diastole phase (in Mmhg).
Cholesterol	This is a type of fat found in the blood normally expressed in mg/dl. But here, we have only levels: 1 corresponding to the lowest (so-called normal) amount and 3 to the highest (so-called very high).
Glucose	It is the level of Glucose in the blood, expressed also in level 1 to 3.
Smoke	It is a binary variable that indicates whether the individual smokes or not (Non-smoking = 0 or Smoking = 1).
Alcool	It is a binary variable that indicates whether the patient is an alcoholic or not (Non-alcoholic = 0 or Alcoholic = 1).
Active	It is a binary variable that indicates whether the patient is physically active or not (No = 0 or Yes = 1).
Cardio	It is a binary variable that indicates whether an patient has cardiovascular disease (CVD) or not (Has no CVD = 0 Has a CVD = 1).
BMI	This is the body mass index. It is classified according to 4 levels: less than 18.5, between 18.5 and 24, between 25 and 29, then greater than or equal to 30.
BPC	This is blood pressure control. It is categorized into 5 levels: to inf 120mmgh to sup at 180mmgh for Systolic. To inf 80mmgh to sup 12mmgh for Diastolic.

## 2 Data Visualization

FIGURE 1: Distribution of age.

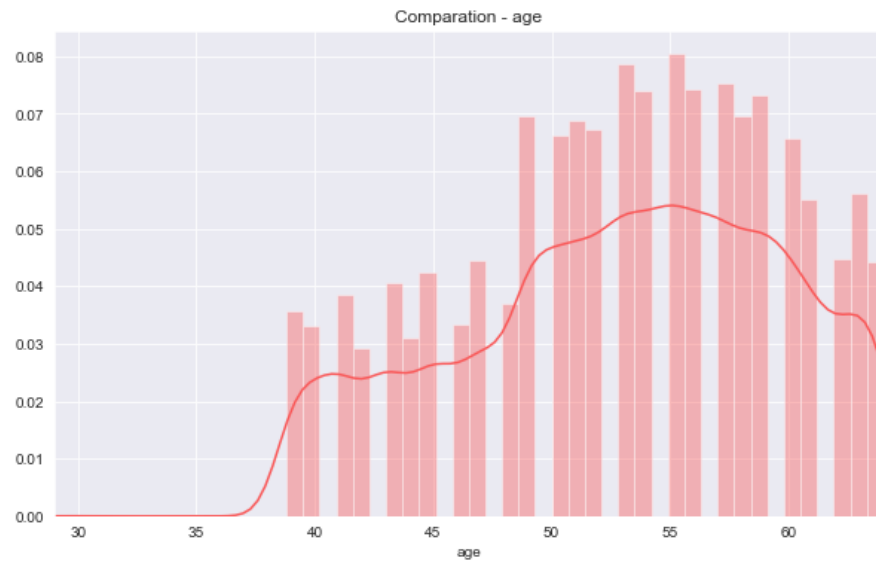


FIGURE 2: Cardiovascular diseases by age.

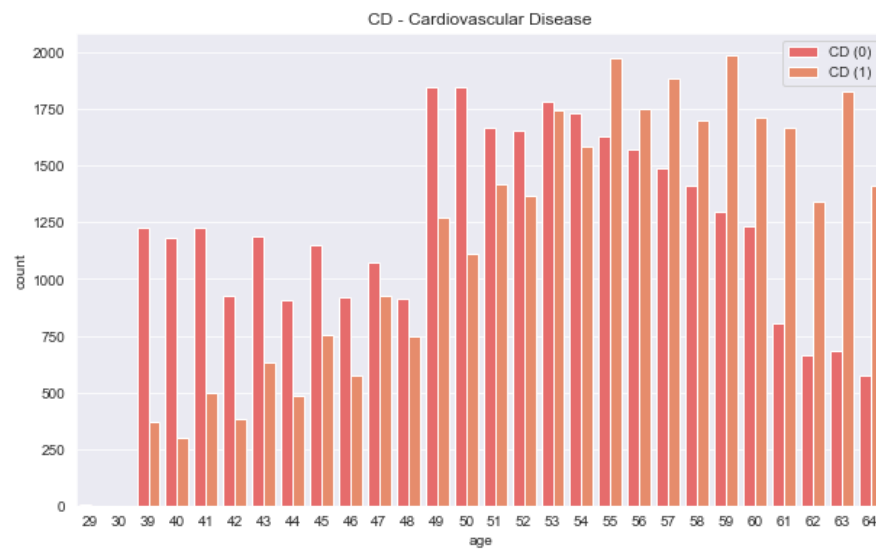
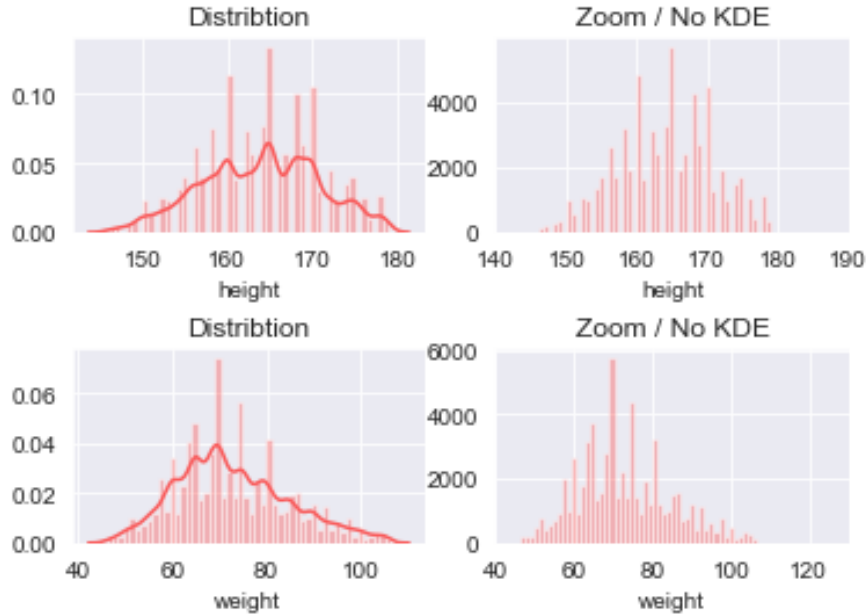
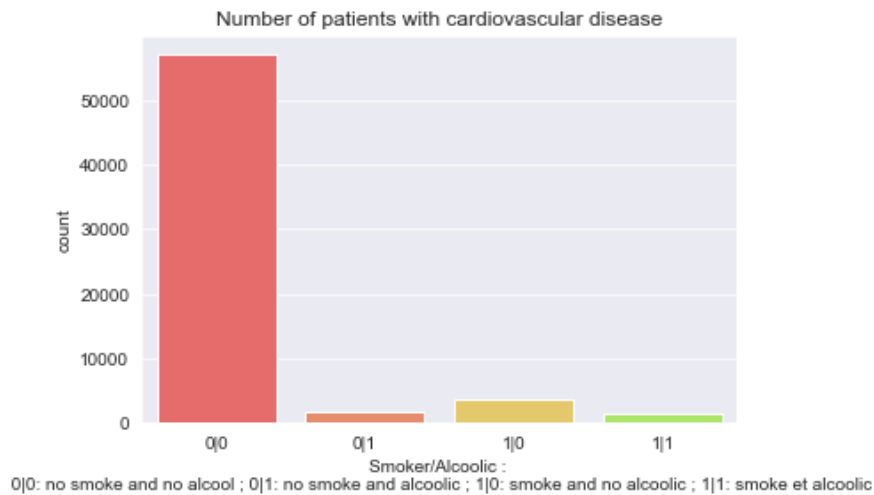


FIGURE 3: Distribution of height and weight.



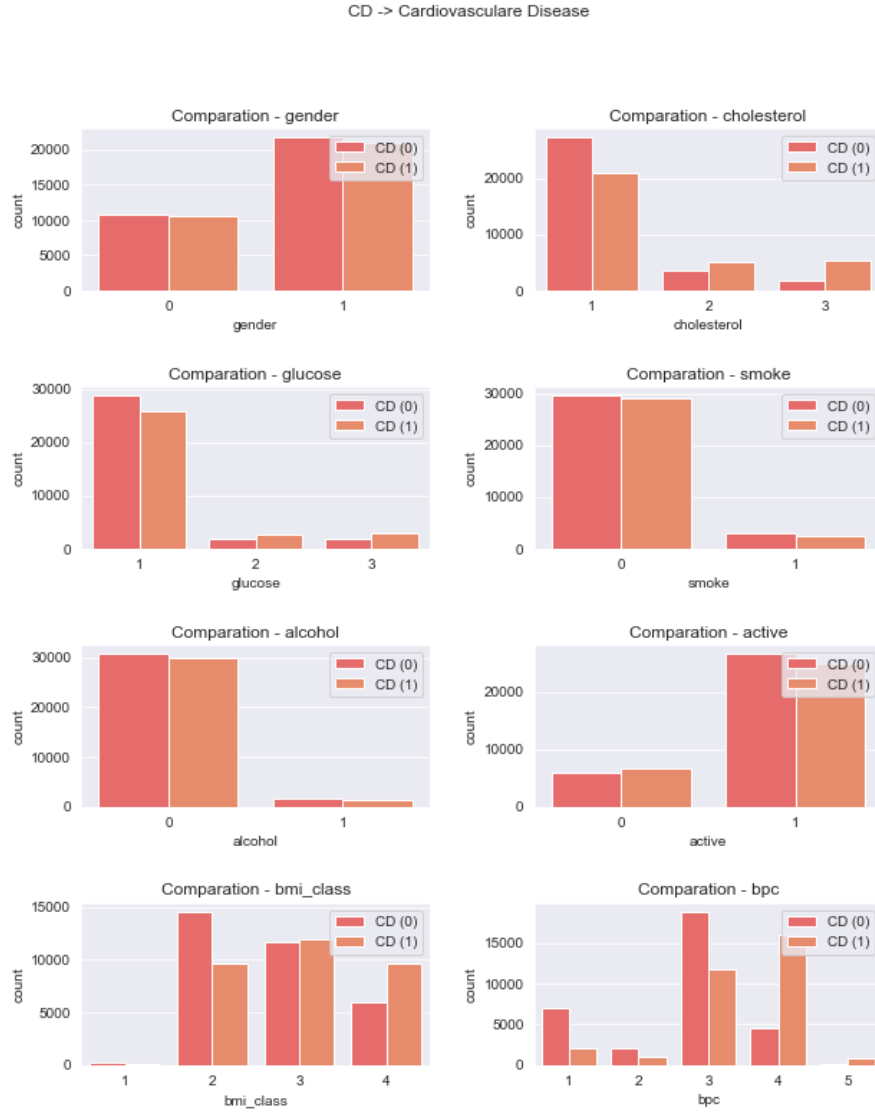
Individuals are between 29 and 65 years of age, most are over 50 years of age. It is clear that as the age of the patient increases, the number of patients with cardiovascular disease increases. Their height is between 130 cm and 207 cm. The weights of individuals are concentrated between 30 and 200 kg. Many patients had inconsistent sizes and weights. So we looked at them as outliers to be removed. In order to be objective and to keep a dataset representative of the population, we have deleted the values of weights and sizes being below the quantile 0.05 and above the quantile 0.975.

FIGURE 4: What about the interaction between Smoke and Alcool.



In our data, we have twice as many women as men. However, there are almost

FIGURE 5: Complete visualization of the dataset.



as many people with cardiovascular disease as there are people without cardiovascular disease, regardless of the patient's gender. Since we notice that there are approximately as many women as men who have cardiovascular disease. Cholesterol is divided into three categories. Level 1 is a normal level. Level 2 is above average and level 3 is very extreme. A large majority of individuals (approximately 50,000) have a normal rate, approximately 9,000 individuals have a high rate, and approximately 7,800 individuals have a very high rate. So overall, the majority of our patients here have normal cholesterol levels. Glucose is divided into three categories in the same way as cholesterol. There are about 58,000 individuals with a normal rate, about 5,000 individuals with a high rate, and about 5,000 individuals with a very high rate. We find the same order of magnitude as for cholesterol. In addition, individuals with high and very high levels of glucose and cholesterol are more affected by cardiovascular disease than those with normal levels. Individuals

who participate in sports are slightly less affected by cardiovascular disease than those who are sedentary. We can see graphically that smoking and drinking alcohol do not seem to illustrate a clear correlation with cardiovascular disease. There are about as many people who don't smoke and don't drink who are affected by cardiovascular disease. This result does not seem real to us, because smoking and drinking are risk factors that increase the possibility of developing cardiovascular disease.

Finally, let's look at the arterial pressures and the size of the patients. Individuals with high systolic and diastolic pressure levels are more affected by cardiovascular disease. Older people are more at risk. On the other hand, individuals with above-normal BMI are slightly more affected by cardiovascular disease.

This result does not seem real to us, because smoking and drinking are risk factors that increase the possibility of developing cardiovascular disease. This result is possibly related to the low representation of these individuals in the data and the measurement of these variables. Since these are binary, they do not give information on the amount smoked or drunk. However, this quantitative information seems important because it is assumed that the greater the quantity, the greater the risk of developing a disease.

## 3 Data Modeling

### 3.1 Tests

Before we start writing an econometric model of its interpretation, we will make assumptions by formulating different hypotheses, sometimes intuitive and other findings observed during the data visualization, that we are going to test in order to find the possible interactions, dependencies between the variables, which must be included in our model.

Hypothesis 1: Are the cholesterol and BMI variables independent?

TABLE 1: Bartlett test of homogeneity of variances.

Data	BMI by Cholesterol
	Bartlett's K-squared = 504.3, df = 2, p-value < 2.2e-16
Data	BMI and Cholesterol
	F = 917.27, df = 2, p-value < 2.2e-16

The cholesterol variable is qualitative while the BMI variable is quantitative, an Anova with 1 factor is used to test the independence between these two variables. In this Anova, we must specify the parameter of homogeneity of variances, for this we perform a Bartlett test which rejects the null hypothesis of equality of variance.



Anova is then performed, whose null hypothesis translates that the average BMI of people is the same in each cholesterol category. Since the p-value is  $2.2 \times 10^{-16}$ , the null hypothesis is rejected and it is concluded that there is a difference in average BMI in the different cholesterol categories. It can also be noticed with the help of boxplots. Thus, the cholesterol and BMI variables are not independent and their interaction will be included in our model.

Hypothesis 2: Are diastolic and systolic variables correlated?

TABLE 2: Pearson's correlation test between Systolic and Diastolic pressure.

Data	Ap hi and Ap lo
	Correlation ( $x=Ap\ hi, y=Ap\ lo$ ) = 0.7

The two variables are quantitative and we can then know the correlation between the two variables. The correlation coefficient between the two variables is 0.7, indicating a strong correlation between the two variables. Thus, the interaction term  $systolic * diastolic$  must be included in the model. This is natural since the first corresponds to the blood pressure measured during the phase of the systole, that is to say during the contraction of the heart. It therefore opposes the second, which corresponds to the blood pressure measured during the release phase of the heart, also called diastole.

Non-parametric chi-squared tests between two binary variables

Hypothesis 3: Does gender affect smoking habits?

TABLE 3: Pearson's chi-squared test.

Data	Gender and Smoke
	X-squared = 7 866.1, df = 1, p-value < $2.2 \times 10^{-16}$

This hypothesis was made following an observation in the data, it is really specific to our data. The variables are qualitative and the non-parametric test of chi-squared is used. Here, the null hypothesis of independence between the two variables is rejected. Thus, the gender and smoke variables are not independent and their interaction will be included in our model.

Hypothesis 4: Are the cholesterol and glucose variables independent?

TABLE 4: Pearson's chi-squared test.

Data	Cholesterol and Glucose
X-squared	= 21 255, df = 4, p-value < 2.2e-16

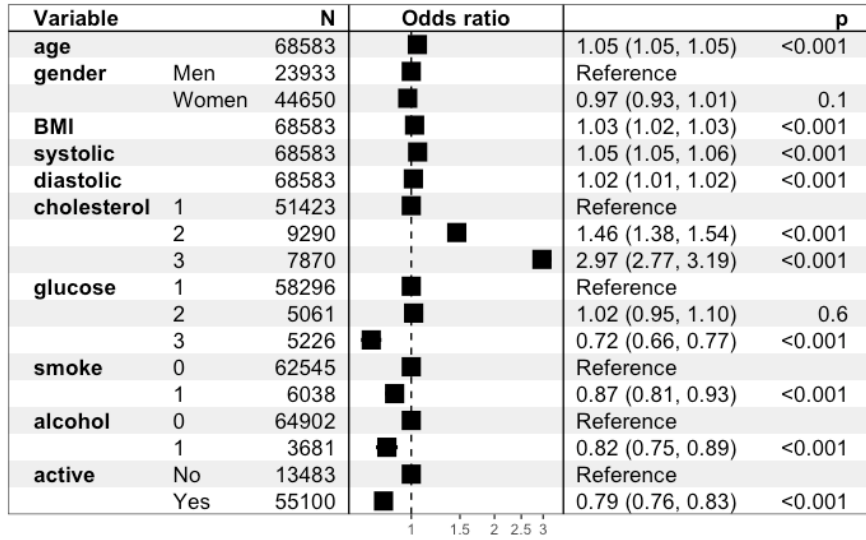
As for the latter hypothesis, the two variables are qualitative and the null hypothesis of independence between the two variables is rejected. Thus, the cholesterol and glucose variables are not independent and their interaction will also be included in our model.

### 3.2 Regressions

The first step is to analyse the model without interaction terms

$$Cardio_i = \beta_0 + \beta_1 Age_i + \beta_2 Gender_i + \beta_3 BMI_i + \beta_4 Systolic_i + \beta_5 Diastolic_i + \beta_6 Cholesterol_i + \beta_7 Glucose_i + \beta_8 Smoke_i + \beta_9 Alcohol_i + \beta_{10} Active_i + \epsilon_i \quad (1)$$

FIGURE 6: Odds ratio graph : first regression result.



$$AIC = 76\,960$$

Akaike's information criterion is a measure of the quality of a statistical model. The goal is to have the most parsimonious model. And therefore to have the lowest AIC.

$$AIC = T \cdot \log\left(\frac{RSS}{T}\right) + 2 \cdot K \quad (2)$$



















Note. RSS is the residual sum of squares. The first term is the fit or approximately the maximized likelihood and  $k$  is the number of regressors in the model.

Here, the reference is  $\text{cardio} = 0$ , meaning not to be affected by cardiovascular disease. Thus, odds ratio less than 1 indicate a decrease in the probability of having cardiovascular disease. Odds ratio greater than 1 indicate an increase in the likelihood of developing cardiovascular disease. Odds ratio equal to 1 indicate no effect. At the 5 percents threshold, all variables are significant with the exception of the Gender and Glucose variables at level 2. Quantitative variables (Age, BMI, Systolic, Diastolic) have odds ratio greater than 1 significant, the more these variables increase the probability of contracting cardiovascular disease increases. It is then noted that the cholesterol variables of levels 2 and 3 are strongly involved in the development of cardiovascular disease. On the other hand, the glucose variables of level 3, Smoke = 1, Alcohol = 1 reduce the risk of developing cardiovascular diseases. This result contradicts what was expected, that is, these variables are risk factors. We suspect that there are problems inherent in our data, particularly on the Smoke and Alcohol variables, whose binary form seems to be very lax, we should take into account the consumption and longevity of these factors. The same observation is made about the lack of precision for the active variable, even though in our model, it reflects what we expected, that is to say that doing sports reduces the risk of cardiovascular disease.

By including terms of interaction, we want to improve the previous model, that is to say, obtain a lower AIC. We include in this new model all the interactions highlighted by the hypothesis tests. We apply a function to this model that allows us to obtain the best model, based on the complete model we propose. This function then looks for a model minimizing the Bayesian criterion.

$$\begin{aligned} Cardio_i = & \beta_0 + \beta_1 Age_i + \beta_2 BMI_i + \beta_3 Systolic_i + \beta_4 Diastolic_i + \beta_5 Cholesterol_i + \\ & \beta_6 Glucose_i + \beta_7 Smoke_i + \beta_8 Alcohol_i + \beta_9 Active_i + \beta_{10} Systolic_i \cdot Diastolic_i \\ & + \beta_{11} BMI_i \cdot Cholesterol_i + \beta_{12} Cholesterol_i \cdot Glucose_i + \epsilon_i \quad (3) \end{aligned}$$

FIGURE 7: Odds ratio graph : second regression result.

Variable		N	Odds ratio		p
<b>age</b>		68583		1.05 (1.05, 1.05)	<0.001
<b>gender</b>	Men	23933		Reference	
	Women	44650		0.97 (0.93, 1.01)	0.1
<b>BMI</b>		68583		1.03 (1.02, 1.03)	<0.001
<b>systolic</b>		68583		1.05 (1.05, 1.06)	<0.001
<b>diastolic</b>		68583		1.02 (1.01, 1.02)	<0.001
<b>cholesterol</b>	1	51423		Reference	
	2	9290		1.46 (1.38, 1.54)	<0.001
	3	7870		2.97 (2.77, 3.19)	<0.001
<b>glucose</b>	1	58296		Reference	
	2	5061		1.02 (0.95, 1.10)	0.6
	3	5226		0.72 (0.66, 0.77)	<0.001
<b>smoke</b>	0	62545		Reference	
	1	6038		0.87 (0.81, 0.93)	<0.001
<b>alcohol</b>	0	64902		Reference	
	1	3681		0.82 (0.75, 0.89)	<0.001
<b>active</b>	No	13483		Reference	
	Yes	55100		0.79 (0.76, 0.83)	<0.001

$$AIC = 76\ 849$$

We note that the AIC has decreased compared to the previous regression without interaction, so the model has a better predictive power. It is observed that the glucose variable has been slightly corrected by the effects of interactions. On the other hand, we may think that the variables  $\text{smoke} = 1$ ,  $\text{alcohol} = 1$  have been slightly corrected, but this impression is due to an extension of the scale of the odds ratio by the right, but we are still not getting the expected results. The interpretation does not change in relation to the model without interaction effects.

### 3.3 Prediction with Machine Learning

In order to answer even more in detail to the question being asked, namely: what are the most important risk factors in the development of these diseases, we are looking at machine learning algorithms to try to predict the fact have a cardiovascular disease or not, and then identify which would be factors to consider. Machine learning is a statistical method for modeling, predicting and solving. There are predictive modeling algorithms to choose from. We need to understand the pros and cons of each to make our choice. The aim is to identify the relationship between the variable to be explained (cardiovascular disease or not) and the explanatory variables (Gender, Age, Cholesterol, BPC, etc.). The algorithm chosen here is based on an evaluation of the different models. By ranking the results classify of all models we can choose the best one for our answer to our problem.

- Logistic Regression is a fairly simple Learning Machine model, it is a regression between a variable to be explained and one or more explanatory variables. The algorithm estimates the probabilities using a logistic function, which is the cumulative logistic distribution.

- The algorithm Naive Bayes classifiers, is a simple classifier based on the application of the Bayes theorem, it also estimates probabilities. Nevertheless, it requires strong assumptions, hence its name Naive and suffers from a learning problem.
- In k-Nearest Neighbors regression, the result is the value for this object. This value is the average of the values of the k closest neighbors.
- The Stochastic Gradient Descent is an iterative method of optimizing an objective function with appropriate smoothing properties. It can be considered as a stochastic approximation of the optimization of the gradient descent, since it replaces the real gradient by an estimate of it.
- The model Random Forest is one of the most popular in Machine Learning. This is an ensemble learning method for classification, regression and other modeling problems. The model operate by constructing a multitude of decision trees (n estimators=100) at training time and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees. Here, we choose n estimator = 100, that is to say that the algorithm constructing a forest with 100 random trees to training the model and predict the output.
- Decision Tree classifier is a method quite similar to Random Forest because it also creates a forest of trees in decision nodes. The branches represent full regressions and the leaves represent target values.
- Support Vector Machines are non-probabilistic linear binary classifiers. In other words, they are supervised learning models with associated learning algorithms that analyze the data used for classification and regression analysis.

When making a prediction model, it is important to focus on F-measures. In statistical analysis, including classification, the F-score (equation 4) is a measure of the accuracy of a test. Its calculation is based on the precision (equation 5) and recall (equation 6) of our algorithm, where accuracy is the number of correctly identified positive results divided by the number of all positive results including those that are not correctly identified, and recall is the number of correctly identified positive results divided by the number of all samples that should have been identified as positive. The model accuracy (equation 7) is the measurement used to determine which model is best at identifying relationships between the variable to be explained and the explanatory variables in a dataset based on the input, or training, data.

$$F1 \text{ score} = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall} \quad (4)$$

$$Precision = \frac{True\ Positive}{True\ Positive + False\ Positive} \quad (5)$$

$$Recall = \frac{True\ Positive}{True\ Positive + False\ Negative} \quad (6)$$

$$Accuracy = \frac{True\ Positive + True\ Negative}{All\ Samples} \quad (7)$$

Models Evaluation	
	Mean Accuracy (%)
Random Forest	73,33
Linear SVC	71,90
k-Nearest Neighbors	71,23
Gaussian Naive Bayes	71,23
Logistic Regression	71,06
Decision Tree Classifier Bayes	62,92
Stochastic Gradient Descent	50,88

FIGURE 8: Receiver Operating Characteristic Curve of Random Forest model.

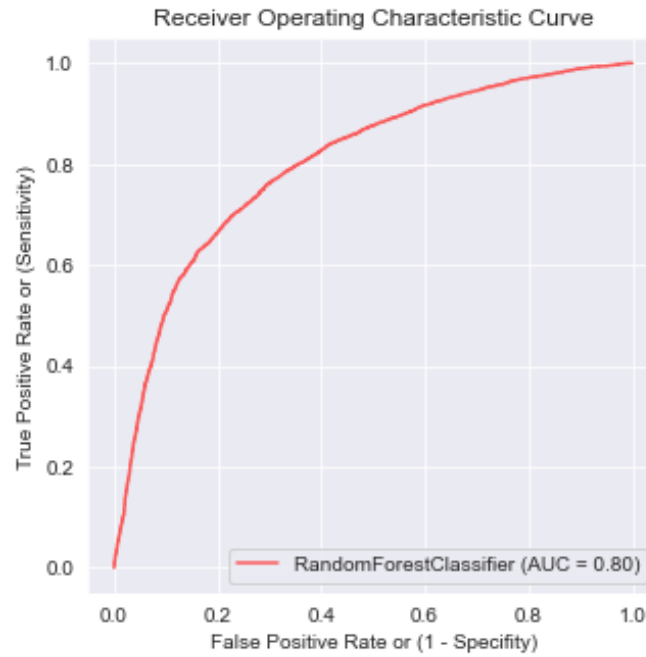


FIGURE 9: Features importance.

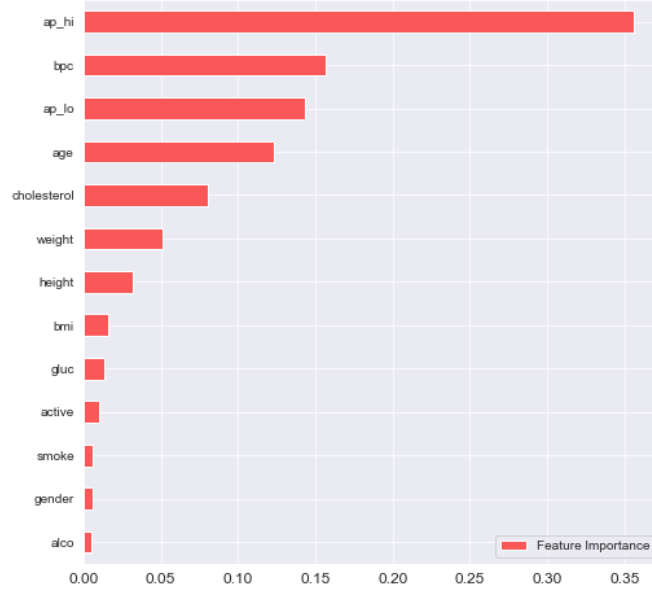


TABLE 5: Random Forest Classifier model

		<i>Precision</i>	<i>Recall</i>	<i>F1-score</i>	<i>Support</i>
Cardiovascular diseases	0	0.71	0.81	0.76	6,599
	1	0.76	0.66	0.71	6,212

	Confusion matrix	
	Positive	Negative
Positive	4,089	1,282
Negative	2,123	5,317

By performing a random forest of 100 trees on 20 percents of the initial sample, our model has an average accuracy of 73.33 percents. That is, our model predicts whether or not to have cardiovascular disease in 73.33 percents of cases. The precision, the probability that a new patient actually being positive given that it is predicted as a positive, of our model is 71 percents for  $CVD = 0$  and 76 percents for  $CVD = 1$ , which means that the model slightly predicts having cardiovascular disease. Recall is rather good for  $CVD = 0$  and less good for  $CVD = 1$ . This is the probability that the patient will be classified as positive ( $CVD = 1$ ), given that it is indeed positive. These are the real positives on the predicted results. The F1-score is also quite good since it is the average of the precision and the recall. The ROC curve confirms our results, with an area below the 0.8 curve.

## 4 Economic Analysis

Once the risk factors to be considered had been demonstrated, it was necessary to determine how prevention should be carried out. We will seek to determine the Cost Incremental Effectiveness Ratio (ICER) through the cost of different treatments but also through the percentages of risk of patients developing these diseases. The aim is to specify how the prevention of these diseases can be carried out effectively.

$$ICER = \frac{effectif \cdot cost}{number\ of\ QALYs\ won} \quad (8)$$

To implement our medical-economic analysis, We relied on two medical research papers: Lisa A. Prosser, Aaron A. Stinnett, Paula A. Goldman, Lawrence W. Williams, Maria G.M. Hunink, Lee Goldman, and Milton C. Weinstein, *Cost-Effectiveness of Cholesterol-Lowering Therapies according to Selected Patient Characteristics*. 2000. Annals of Internal Medicine, Vol 132, pages 769-779. And C. Petite C. A. Meier, *Cholestérol : qui ne faut-il pas traiter ?*. 2002. Revue Médicale Suisse 2002, Vol -2. 22426. We created a new fully randomized dataset composed of 15 percents of the initial data observations. We then isolated men and women and then calculated the Framingham score (associated with each patient in this randomized dataset). Many versions of this score exist, but the theoretical framework remains the same: it is an estimate of the possibility of cardiovascular disease in the next ten years in a patient. The more points a patient has, the higher the risk at 10 years of age. We adapted it to our data so that we could calculate it optimally. The goal is to assign a number of points to each patient according to different variables: age, cholesterol level, whether you are a smoker or not, blood pressure. For example, we have implemented two prevention strategies: a primary prevention for patients who have never had cardiovascular disease – RRC = 0 in our database – and a secondary prevention for patients with cardiovascular disease. Vascular – CVD = 1 –.

We calculated the costs of two treatments for primary prevention. The first treatment is diet, the goal of this treatment is to have a healthier and more varied diet. The second treatment is the use of Statin, it is a drug that helps to stop the obstruction of the blood vessels but also prevents the aggravation of the obstruction already present, this drug will be prescribed in low dose. Secondary prevention is reserved for those who have already developed cardiovascular diseases in order to reduce their aggravation with the help of a higher dose of Statin. For all these treatments, and regardless of prevention, there are ancillary costs that increase the amount of treatment over the next 10 years: visits to the doctor and analyses to improve the follow-up of the patient. On the other hand, we will not take into account here the cost differential that may exist between women and men in these ancillary costs. We'll make it look like they were the same at given sex.



TABLE 6: Costs of treatments.

Primary prevention	Treatment 1: Diet	108 USD per year
	Treatment 2: Statin	1,512 USD 1st year, after 1,318 USD per year
Secondary prevention	Treatment 1: Statin	1,329 USD per year

*Note. The cost of the diet is rather low, because it represents in reality only the follow-up of a doctor and different analyses every year. The cost of Statin is determined by the average selling price of the drug. There are also additional ancillary costs as follow-up requires more medical visits and testing each year.*

Once we have defined the costs, we need to determine the number of Quality-Adjusted Life Year (QALYs), “quality-weighted life year”, an economic indicator to estimate the value of life, which is essential to calculate the ICER.

Based on Paul S. Chan, Brahmajee K. Nallamothu, Hitinder S. Gurm, Rodney A. Hayward, and Sandeep Vijan, *Incremental Benefit and Cost-Effectiveness of High-Dose Statin Therapy in High-Risk Patients With Coronary Artery Disease*. 2007. *Circulation*, Vol 115, Issue 18. And Adam Gilden Tsai, Henry A. Glick, David Shera, Linda Stern, Frederick F. Samaha, *Cost-Effectiveness of a Low-Carbohydrate Diet and a Standard Diet in Severe Obesity*. 2005. *Obesity Society*, Vol 13, issue 10, we defined for primary prevention patients – CVD = 0 – the QALYs gain of treatment 2 is defined as 0.35 QALYs. For those in secondary prevention – with CVD = 1 –, a gain of 0.10 QALYs is observed with Statin therapy. A rather small gain in QALY especially for patients already suffering from the disease since statin is a drug that does not cure but that prevents the disease from developing. In addition it provides many side effects. The second article does not give us an exact number of QALYs gained with a diet treatment, but only an interval between 0.29 and 1 QALYs. Since this treatment applies to patients with only possible risks of occurrence of these diseases, we will choose to set, subjectively, 0.645 the QALYs gain of this treatment.

FIGURE 10: Primary prevention: Treatment 1, Diet, Men, CVD = 0.

		Cost in Dollar per QALYS (\$)					
		34-39y	40-44y	45-49y	50-54y	55-59y	60-64y
Risk at 10years							
	≤ 5%	125,581	510,698	532,465	555,907	406,884	0
	>5%	0	0	21,767	170,791	197,581	405,209

FIGURE 11: Primary prevention: Treatment 2, Statin, Men, CVD = 0.

Cost in Dollar per QALYS (\$)						
	34-39y	40-44y	45-49y	50-54y	55-59y	60-64y
Risk at 10years						
≤ 5%	2,865,857	11,654,486	12,151,234	12,686,194	9,285,377	0
>5%	0	0	496,749	3,897,566	4,508,949	9,247,166

FIGURE 12: Primary prevention: Treatment 1, Diet, Women, CVD = 0.

Cost in Dollar per QALYS (\$)						
	34-39y	40-44y	45-49y	50-54y	55-59y	60-64y
Risk at 10years						
≤ 5%	182,512	915,907	982,884	1,600,744	1,229,023	750,140
>5%	0	0	0	1,674	1,674	0

FIGURE 13: Primary prevention: Treatment 2, Statin, Women, CVD = 0.

Cost in Dollar per QALYS (\$)						
	34-39y	40-44y	45-49y	50-54y	55-59y	60-64y
Risk at 10years						
≤ 5%	4,165,046	20,901,651	22,430,109	36,530,126	28,047,189	17,118,720
>5%	0	0	0	38,211	38,211	0

First, concerning men, we notice (looking at the tables of the step 1) that a large majority of them have a score of less than or equal to 5 percents, when they have no cardiovascular diseases. Older adults, on the other hand, have risk percentages greater than 5 percents. In addition, we calculated the risk of cardiovascular diseases for step 2 patients, that is to say those already being affected by cardiovascular disease. Thus, the results of step 2 will serve us to decision criteria for deciding which prevention to implement.

FIGURE 14: Secondary prevention: Treatment 1, Statin, Women, CVD = 1.

Cost in Dollar per QALYS (\$)						
	34-39y	40-44y	45-49y	50-54y	55-59y	60-64y
Risk at 10years						
≤ 5%	3,322,500	16,745,400	24,985,200	30,965,700	6,910,800	0
>5%	265,800	0	12,226,800	21,131,100	63,526,200	61,532,700

FIGURE 15: Secondary prevention: Treatment 1, Statin, Men, CVD = 1.

Cost in Dollar per QALYS (\$)						
	34-39y	40-44y	45-49y	50-54y	55-59y	60-64y
Risk at 10years						
≤ 5%	6,379,200	28,839,300	53,356,100	101,801,400	135,159,300	113,895,300
>5%	0	0	0	265,800	1,993,500	1,727,700

Looking at the results of step 2, we notice that most of the men with cardiovascular disease have risk percentages less than or equal to 5 percents. While very few have scores above 5 percents. These results confirm the findings at the end of step 1. Thus, prevention can be justified mainly on the men with risk percentages less than or equal to 5 percents. These patients should be Diet, which is the most cost-effective treatment. Indeed, he can put himself in place for a lower cost than treatment with Statin and brings more QALYs to the patients. For older patients, as there are very few of them, it is possible to they should be treated with Statin. Therefore, it is recommended to treat men from the age of 40 by combining the two treatments, the Diet for almost all patients and Statin only for the most risky and elderly patients.

For women, the interpretation is a little different and is not intuitive. We found that women without cardiovascular disease (step 1) have scores very much less than or equal to 5 percents. In contrast, those who have developed a disease (step 2) have, on average, a score above 5 percents. This would mean that prevention must be focus first on women who score above 5 percents when they have not no cardiovascular diseases. Indeed, this is what we recommend because if their score is high, these women have a very good chance of developing cardiovascular disease. So that's towards the latter, which prevention must turn to.

## 5 Conclusion

After the fact, it is possible to answer several questions. First, we can affirm that the prevention of cardiovascular disease is essential in order to be cost-effective for states. Prevention must be adapted according to several factors: these are: in patients, as a priority blood pressure (systolic and diastolic), the level of cholesterol in the blood but also the age of the patients. We could see, by the cost modelling by QALYs, that prevention should be directed to older when the patient is male and towards younger patients when considering women. Overall, many other factors contribute to an increased risk of the onset of a cardiovascular disease such as smoking or not smoking, or physical and sports activity, or glucose levels of patients. So many factors to be monitored to effectively and sustainably prevent the development of these diseases. Last part of our duty is to show that the most cost-effective treatment is a healthy diet, to reduce any risk of developing a disease when has never contracted any. Randomizing our sample gives us a additional robustness to really show that prevention needs to be different according to sex. Statin step 2 treatment for women appears as the least cost-effective because they are few, compared to men, in develop cardiovascular diseases. Concerning men, the results are rather We have a mirror effect with the 10-year risk level observed. We conclude that for patients with a risk less than or equal to 5 percents, Diet treatment is the most cost-effective. On the other hand, many older men have relatively high may require treatment with Statin at a low dose. In addition, our results are reinforced by the fact that if new datasets are generated on the basis of the principal, the cost in USD per QALYs, the proportion of women and men with and without diseases cardiovascular will be approximately the same, thus our results and conclusions will be the same.

## 6 Discussion

We still have to qualify our comments, first of all, with respect to the veracity of our data. They were found on the *Kaggle website* and therefore did not come from an official source. Second, the measurement of some variables is wrong. For example, the Smoke and Alcohol and Active variables are binary. Thus, they provide only yes or no information, but tell us nothing about the level of consumption (of practice for the active variable). An individual who smokes slightly will have a higher percentage risk of developing cardiovascular disease than an individual who smokes occasionally but is counted in the “yes”. The same applies to patient who drink alcohol. So this goes back to our questions about the calculation of the Framingham score, the allocation of points should be different according to the levels of tobacco consumption, and thus, the percentages of risk at 10 years modified. Finally, our database seems to present many problems, both in the measurement of its variables, and in its non-representativeness of a lambda population.

## References

Paul S. Chan, Brahmajee K. Nallamothu, Hitinder S. Gurm, Rodney A. Hayward & Sandeep Vijan. (2007). *Incremental Benefit and Cost-Effectiveness of High-Dose Statin Therapy in High-Risk Patients With Coronary Artery Disease*. CirculationNaha, Vol 115, Issue 18.

Adam Gilden Tsai, Henry A. Glick, David Shera, Linda Stern & Frederick F. Samaha. (2005). *Cost-Effectiveness of a Low-Carbohydrate Diet and a Standard Diet in Severe Obesity*. Obesity Society, Vol 13, issue 10.

Hélène Hubert-Yahi, Maître de Conférences en Sciences Economiques. Cours d'Econométrie de la santé. 2020. Master 1 Econométrie – Statistiques (MoSEF), Université Paris 1 Panthéon-Sorbonne.

C. Petite C. & A. Meier. (2002). *Cholestérol : qui ne faut-il pas traiter ?*. Vol -2. 22426.

Lisa A. Prosser, Aaron A. Stinnett, Paula A. Goldman, Lawrence W. Williams, Maria G.M. Hunink, Lee Goldman & Milton C. Weinstein. (2000). *Cost-Effectiveness of Cholesterol-Lowering Therapies according to Selected Patient Characteristics* . Annals of Internal Medicine, Vol 132, pages 769-779.

# Annexes

FIGURE 16: Calculation method used to obtain the Framingham score.

Estimate of 10-Year Risk for Men (Framingham Points Scores)						Estimate of 10-Year Risk for Women (Framingham Points Scores)					
Age		Points				Age		Points			
20-34		-9				20-34		-7			
35-39		-4				35-39		-4			
40-44		0				40-44		0			
45-49		3				45-49		3			
50-54		6				50-54		6			
55-59		8				55-59		8			
60-64		10				60-64		10			
Cholesterol	Age 20-39	Age 40-49	Age 50-59	Age 60-64		Cholesterol	Age 20-39	Age 40-49	Age 50-59	Age 60-64	
Niveau 1	0	0	0	0		Niveau 1	0	0	0	0	
Niveau 2	4	3	2	1		Niveau 2	4	3	2	1	
Niveau 3	7	5	3	1		Niveau 3	8	6	4	2	
	Age 20-39	Age 40-49	Age 50-59	Age 60-64			Age 20-39	Age 40-49	Age 50-59	Age 60-64	
No Smoke	0	0	0	0		No Smoke	0	0	0	0	
Smoke	8	5	3	1		Smoke	9	7	4	2	
BPC	Niveau 1	Niveau 2	Niveau 3	Niveau 4	Niveau 5	BPC	Niveau 1	Niveau 2	Niveau 3	Niveau 4	Niveau 5
CVD = 0	0	0	1	1	2	CVD = 0	0	1	2	4	4
CVD = 1	0	1	2	2	4	CVD = 1	0	3	4	5	6
Point Total	10-Year Risk %					Point Total	10-Year Risk %				
< 0	1					< 9	1				
1	1					9	1				
2	1					10	1				
3	1					11	1				
4	1					12	1				
5	2					13	1				
6	2					14	2				
7	3					15	3				
8	4					16	4				
9	5					17	5				
10	6					18	6				
11	8					19	8				
12	10					20	11				
13	12					21	14				
14	16					22	17				
15	20					23	22				